

## Review

# Trigger Tool–Based Automated Adverse Event Detection in Electronic Health Records: Systematic Review

Sarah N Musy<sup>1,2</sup>, MSc; Dietmar Ausserhofer<sup>1,3</sup>, RN, PhD; René Schwendimann<sup>1,4</sup>, RN, PhD; Hans Ulrich Rothen<sup>5</sup>, MD, PhD; Marie-Madlen Jeitziner<sup>5</sup>, RN, PhD; Anne WS Rutjes<sup>6,7</sup>, PhD; Michael Simon<sup>1,2</sup>, RN, PhD

<sup>1</sup>Institute of Nursing Science, University of Basel, Basel, Switzerland

<sup>2</sup>Nursing & Midwifery Research Unit, Inselspital Bern University Hospital, Bern, Switzerland

<sup>3</sup>College for Health Care Professions, Claudiana, Bolzano, Italy

<sup>4</sup>University Hospital Basel, Patient Safety Office, Basel, Switzerland

<sup>5</sup>Department of Intensive Care Medicine, Inselspital Bern University Hospital, Bern, Switzerland

<sup>6</sup>Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland

<sup>7</sup>Institute of Primary Health Care (BIHAM), University of Bern, Bern, Switzerland

**Corresponding Author:**

Michael Simon, RN, PhD  
Institute of Nursing Science  
University of Basel  
Bernoullistrasse 28  
Basel, 4057  
Switzerland  
Phone: 41 61 267 09 12  
Email: [m.simon@unibas.ch](mailto:m.simon@unibas.ch)

## Abstract

**Background:** Adverse events in health care entail substantial burdens to health care systems, institutions, and patients. Retrospective trigger tools are often manually applied to detect AEs, although automated approaches using electronic health records may offer real-time adverse event detection, allowing timely corrective interventions.

**Objective:** The aim of this systematic review was to describe current study methods and challenges regarding the use of automatic trigger tool-based adverse event detection methods in electronic health records. In addition, we aimed to appraise the applied studies' designs and to synthesize estimates of adverse event prevalence and diagnostic test accuracy of automatic detection methods using manual trigger tool as a reference standard.

**Methods:** PubMed, EMBASE, CINAHL, and the Cochrane Library were queried. We included observational studies, applying trigger tools in acute care settings, and excluded studies using nonhospital and outpatient settings. Eligible articles were divided into *diagnostic test accuracy studies* and *prevalence studies*. We derived the study prevalence and estimates for the positive predictive value. We assessed bias risks and applicability concerns using Quality Assessment tool for Diagnostic Accuracy Studies-2 (QUADAS-2) for diagnostic test accuracy studies and an in-house developed tool for prevalence studies.

**Results:** A total of 11 studies met all criteria: 2 concerned *diagnostic test accuracy* and 9 *prevalence*. We judged several studies to be at high bias risks for their automated detection method, definition of outcomes, and type of statistical analyses. Across all the 11 studies, adverse event prevalence ranged from 0% to 17.9%, with a median of 0.8%. The positive predictive value of all triggers to detect adverse events ranged from 0% to 100% across studies, with a median of 40%. Some triggers had wide ranging positive predictive value values: (1) in 6 studies, hypoglycemia had a positive predictive value ranging from 15.8% to 60%; (2) in 5 studies, naloxone had a positive predictive value ranging from 20% to 91%; (3) in 4 studies, flumazenil had a positive predictive value ranging from 38.9% to 83.3%; and (4) in 4 studies, protamine had a positive predictive value ranging from 0% to 60%. We were unable to determine the adverse event prevalence, positive predictive value, preventability, and severity in 40.4%, 10.5%, 71.1%, and 68.4% of the studies, respectively. These studies did not report the overall number of records analyzed, triggers, or adverse events; or the studies did not conduct the analysis.

**Conclusions:** We observed broad interstudy variation in reported adverse event prevalence and positive predictive value. The lack of sufficiently described methods led to difficulties regarding interpretation. To improve quality, we see the need for a set of recommendations to endorse optimal use of research designs and adequate reporting of future adverse event detection studies.

**KEYWORDS**

patient safety; electronic health records; patient harm; review, systematic

## Introduction

In recent decades, patient safety and quality of care have become a top priority in health care [1-3]. This has led to significant progress, especially regarding innovative use of electronic health records (EHRs). Adverse events (AEs), injuries attributed to medical care that are independent of the patient's underlying condition, nevertheless remain a persistent problem. Apart from the impact on patients, they entail large human and financial burdens at every health care system level [4]. Regarding patient health, AEs' negative consequences include extended hospital stays, higher readmission rates, and higher mortality [5]. Furthermore, AEs may lead to the patients' and their families' loss of trust in their health care professionals (HCPs), their health care system, or both [3]. The estimated prevalence of AEs in hospital inpatients ranges from 3% to 40% in acute care settings [2,6-10]. The wide range reflects the challenges involved in detecting and tracking AEs accurately [11].

To improve patient safety, health care organizations need valid and reliable tools to detect and assess AEs [12]. Several tools exist, but their ability to identify AEs is limited, and none of them are broadly accepted [13-15]. Currently, enumerating specific events that endanger patients depends mainly on voluntarily reporting by health care staff [16,17]. Systematic evaluations of this approach showed endemic underreporting, with only 2% to 8% of all harmful events being identified [18-20]. To depict the situation more robustly, the US Agency for Healthcare Research and Quality (AHRQ) published a set of Patient Safety Indicators (PSIs). Using administrative datasets, PSIs identify potential AEs, but are highly susceptible to variations in coding practice and are limited by many outcomes being easily concealed in the medical record [14]. Therefore, they miss a substantial fraction of AEs (low sensitivity), while producing a substantial fraction of false positive results (low specificity) [13,21].

One promising method is the Global Trigger Tool (GTT), developed by the Institute for Healthcare Improvement (IHI) [22]. Providing a structured method for identifying AEs from patient records [23,24], the GTT is a retrospective record review instrument that uses a list of triggers (or clues), ie, data elements within the health record, to alert reviewers to the potential presence of AEs [22,25]. By focusing on triggers within patient records, the GTT has demonstrated to identify up to ten times as many AEs as other detection methods [13]. Various studies have used the GTT, where some modified the methods, eg, by modifying the set of triggers, or by modifying the review process (eg, one reviewer instead of two for trigger identification). We refer to these modified versions as *trigger tool methodology*, reserving the term *GTT methodology* for the IHI's original procedures [24].

The trigger tool was developed as a manual approach, ie, for application by HCPs reviewing patient records. Recently, an

increasing interest developed for semi or fully automated AE detection methods using EHRs where lesser time and personnel resources are required for the AE identification [25-28]. *Prospective AE detection* would supply real-time feedback to HCPs, allowing timely interventions. The development of automated surveillance systems using EHR data has greatly facilitated AEs' identification [28].

Semi or fully automated AEs detection methods show promise to efficiently measure AEs. Nevertheless, evidence need to be summarized based on the current literature to gather information for future development and implementation in a health care organization. As a variety of AEs' detection methods exist, we decided to focus on trigger tool-based AEs detection methods only, allowing comparisons between studies as suggested in a previous systematic review on automated detection of patient harm [29]. As trigger tool methodology has shown higher sensitivity compared with other detection methods, we considered the manual trigger tool as the *gold standard*. This systematic review aimed to describe current study methods and challenges regarding the use of automatic trigger tool-based AE detection methods in EHRs in acute care settings. In addition, we aimed to appraise the applied studies' designs and to synthesize estimates of AE prevalence and diagnostic test accuracy (DTA) of automatic detection methods using manual trigger tool as a reference standard.

## Methods

### Search Strategy and Study Selection

This systematic review followed the recommendations of the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy [30], the Cochrane Handbook for Systematic Reviews of Interventions [31], and the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines for the reporting of systematic reviews [32].

Hausner et al's approach was applied to develop a robust search strategy ([Multimedia Appendix 1](#)) [33,34]. In PubMed's basic search mode, we entered the following medical subject headings (MeSH) and free-text terms for title and abstract fields: (*trigger OR triggers*) AND (*chart OR charts OR identif\* OR record OR records*) AND (*adverse OR medical errors*). The focus of the search was on "trigger" and not on GTT, as we aimed to include studies using variations of the original GTT. The search strategy was transposed to EMBASE, CINAHL, and the Cochrane Library, and terms were mapped to the appropriate keywords (eg, from MeSH to Emtree). Studies published in English, French, German, Italian, or Spanish were considered without restrictions concerning the years of the publication. In addition to searching the bibliographic databases, the team identified additional relevant literature from most common journals publishing in the field of trigger tool: *BMJ Quality & Safety*, *Journal of Patient Safety*, and *International Journal for Quality in Health*. For pragmatic reasons, the research team decided to

limit the hand search of the most common journals to the years 2014 to 2017. The search was conducted in November 2015 with updates in April 2016 and July 2017.

We included observational studies that applied a trigger-based tool to detect AEs in EHRs in any acute care setting. We defined the target population of interest as patients hospitalized for at least 48 hours for any reason. The evaluated trigger tool approach (index test) had to involve either semi (ie, one part of the process still manual) or fully automatic identification of AEs [29]. Regarding DTA studies, we opted for a reference standard that produced a relatively low rate of missed AEs alongside an acceptable rate of false positive test results (events flagged as AEs that, upon examination, did not qualify as AEs). As shown by Classen et al (2011), manual trigger-based tools met our target criteria [13]. We excluded studies from nonhospital settings (eg, long-term care), outpatient clinics, or that concerned with nonprimary research (eg, systematic reviews or editorials).

The eligible articles were divided into two sets: (1) an automated trigger tool in comparison with a manual trigger tool method for AE detection, potentially enabling the evaluation of the trigger tool's DTA (*diagnostic test accuracy studies*) and (2) application of an automated trigger tool without cross-verification with a manual trigger tool method, enabling us to synthesize the prevalence of AEs and the applied methods. We refer to the latter group as "*prevalence studies*" throughout this paper.

### Data Extraction

Two main reviewers (SNM and MS) each screened half of the retrieved titles and abstracts for relevance according to the criteria outlined above. The other members of the research team each screened a quarter of the retrieved titles and abstracts, allowing double screening for all citations. Full-text screening was independently assessed by the main two reviewers, where disagreements were resolved by discussion, or by consulting the entire research team, if necessary.

Detailed study information was extracted into tables by SNM and a master student as part of her training. We used standardized piloted extraction sheets to tabulate variables related to design, sample population characteristics, applied trigger tool methodology, type and number of reviewers and triggers, and outcome data expressed as AE prevalence (overall and by AE type). To estimate DTA, we used 2x2 tables. Whenever possible, we derived the positive predictive value (PPV) of the triggers used. PPV is calculated by dividing the number of true positive triggers related to confirm AEs by the total number of positive triggers.

Unresolved disagreements or uncertainties between SNM and the master student were discussed and resolved in the research team, which included experienced systematic review and GTT methodologists, clinicians, and nurses, each with more than 10 years of experience in their specialty.

### Quality Assessment

We assessed the risk of bias and the concerns regarding applicability of all included studies. With respect to DTA, we assessed the quality of the included studies with the QUADAS-2 instrument, which we adapted for use as recommended by its authors [35]. The tool consists of four domains: (1) patient selection, (2) index test, (3) reference standard, and (4) flow and timing. Each domain contains signaling questions for risk of bias and concerns regarding applicability except the domain flow and timing. Each signaling question has three answer options: yes, no, and unclear. On the basis of the overall rating of the reviewers, an assessment can be made in each domain concerning bias and applicability. For example, one signaling question in the domain patient selection is, "Was a consecutive or random sample of patients enrolled?" All adaptations of the instrument are explained in [Multimedia Appendix 2](#).

For prevalence studies, the research team generated a new tool based on the structure of the QUADAS-2 instrument to assess the risk of bias and concerns regarding applicability. The new tool consisted of six domains: patient selection, reviewer and algorithm selection, automatic detection method, outcomes, and flow and timing. All but outcomes and flow and timing included a section on risk of bias and concern regarding applicability, including signaling questions ([Multimedia Appendix 3](#)). The goal of this quality assessment was to verify the quality of the semi or fully automated trigger tool studies by focusing specifically on algorithm development and the basis for choosing each trigger.

Quality assessment process was conducted by one reviewer (SNM), and each research team member received at least one study to compare the results with SNM. Members did each task individually; then all results were discussed jointly by the multidisciplinary research team.

### Statistical Analyses

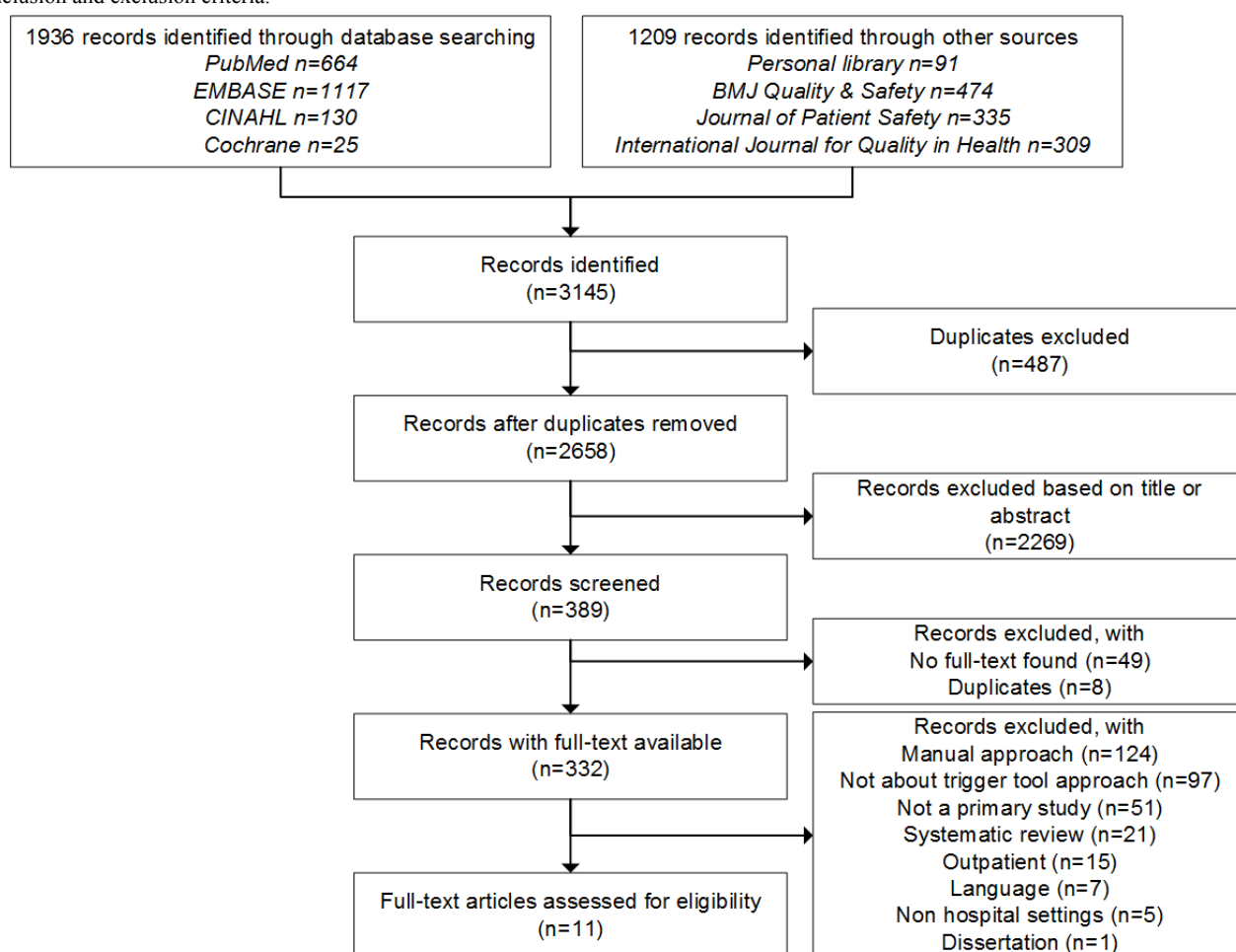
As we anticipated a paucity of evidence on DTA data, we deemed formal meta-analyses not feasible. For the total number of AEs and each type of AE, we present percentages with 95% CIs for prevalence, PPV, and rate of false negative test results. For the prevalence studies, we provided percentages for prevalence and PPV with 95% CIs, for AE overall and per type of trigger.

## Results

### Search Strategy and Study Selection

After removing duplicates, 2658 citations were identified via our search strategy, the core journals, and our personal library. Of these, 11 met all selection criteria: 2 concerned diagnostic test accuracy studies and 9 prevalence studies. A detailed view of the included studies is provided with a flow diagram in [Figure 1](#).

**Figure 1.** Flow diagram of the number of studies found with the search strategy, studies screened, and reasons for exclusions. Eleven studies fulfilled all inclusion and exclusion criteria.



## Description of Included Studies

The studies were published from 2005 to 2016. Of these, 9 were conducted in the United States [25,36–43], 1 in Denmark [44], and 1 in the United Kingdom [45]. Concerning the study populations, 7 were conducted among pediatric or neonatal patients [25,36,37,39,40,43,45], 3 among adults [38,41,42], and 1, published in abstract form only, provided no population information [44]. Two studies conducted multisite trigger tool research [25,42]. The bed capacity of the hospitals involved ranged from 26 to 1000 beds, with a median of 306. One study provided no information on bed capacity [36]. Further details concerning the design and the characteristics are summarized in Table 1.

One study explicitly followed the IHI guidelines for GTT [42], whereas the other studies used different trigger tool approaches or did not specify whether they followed the IHI guidelines. Concerning methodology, 6 studies addressed only adverse drug events (ADEs) [36,38,39,42,43,45], and only 1 used a fully automatic AE detection approach [44]. Definitions used to define and reference AEs varied [36–38,40–43,45] or were absent [25,39,44]. Seven studies used decision or agreement of the reviewer(s) to confirm an AE, without reference to the indexing method used [25,36–40,43]; and 2 studies omitted any explanation of the process [44,45]. Five studies did not consider preventability [38,39,42,44,45]; in the remainder, definitions

varied. Three studies used the definition of a preventable AE as an “event leading to action deviating from the local standard of care” [37,40,43]; one study used a scale from “1—virtually no evidence for management causation” to “6—virtually certain evidence for management causation,” but without describing the applied cutoff [36]. One study used a 6-point confidence scale from “1—virtually no evidence of preventability” to “6—virtually certain evidence of preventability,” with a score  $\geq 4$  denoting preventability [41]; and one study merely described that preventability was determined by the reviewers’ decision [25]. Five studies assessed AEs’ severity using the National Coordinating Council for Medication Error Reporting and Prevention (NCC MERP) categories [25,37,39,40,43]; 4 studies did not state their assessment strategies [38,42,44,45]; one study categorized them according to four levels of severity (life-threatening, serious, clinically significant, or trivial) [41]; and one used NCC MERP categories together with the Common Terminology Criteria for Adverse Events version 4.03, ranging from mild (grade 1) to death (grade 5) [36]. Information concerning the data source, the triggers, and the reviewer(s) are detailed in Table 2.

Concerning the methodology, 5 studies came from the “Automated Adverse Event Detection Collaborative,” which is a consortium to facilitate the use of automated triggers in pediatric hospitals [25,36,37,39,40,43]. They all used the same approach, where an EHR-driven surveillance system was used,



yet not providing detailed information on the software. Every night, trigger reports were automatically generated for laboratory results [37,40,43], medications levels in the blood [36,39], or both [25]. A clinical analyst examined every trigger by reviewing the EHR and interviewing care providers. The potential AEs were reviewed by specialists: pharmacists, physicians, endocrinologists, or anesthesiologists. The clinical analyst met with a multidisciplinary team, the “Automated Adverse Event Detection Steering Committee,” monthly to present the results. Two studies used natural language processing (NLP) to extract information from EHRs [41,44]. NLP is defined as a technique extracting information from narrative text and transforming it into structured text [41,46]. NLP is able to deal

with synonyms, negations, and abbreviations used in narrative text. To build queries, SAS Text Miner tool [44] was used; however, no further details were provided. Structured Query Language [41] was used to identify AEs in the EHRs. The other studies used (1) Electronic trigger-detection messages that were sent automatically to two pharmacist reviewers [42], (2) Computerized trigger alert system that sent an alert to the project manager within 24 hours when conditions defined by the trigger algorithm were detected [38], and (3) electronic algorithms where triggers were identified automatically [45]. No further details concerning the development, the algorithms, or the tools used were given. Description of the methods are explained in Table 3.

**Table 1.** Design and characteristics of the sample and population of the included studies.

Study	Population	Time frame (months)	Sample size	Setting
<b>Diagnostic test accuracy studies</b>				
Gerdes and Hardahl, 2013 [44]	Not stated	26	500	Not stated
O’Leary et al, 2013 [41]	Adults, exclusion of patients admitted under observation status and cared for by either of the two medical record abstractors	12	250	General internal medicine
<b>Prevalence studies</b>				
Call et al, 2014 [36]	Children	48	390	Oncology Hematology
Dickermann et al, 2011 [37]	Children, exclusion weekend days for 5 months because of resource limitations	12	13,526	General internal medicine Surgical care Emergency department Intensive care unit (ICU) Psychiatric unit
Lim et al, 2016 [42]	Adults	3x1 month	Not stated	Not stated
Moore et al, 2009 [38]	Adults	5	456	General internal medicine Surgical care Obstetrics or gynecology
Muething et al, 2010 [39]	Children	21 for one trigger and 16 for another one	Not stated	Not stated
Nwulu et al, 2013 [45]	Not stated	12	54,244	Not stated
Patregnani et al, 2015 [43]	Children	52 for one trigger; 40 and 60 for the others	Not stated	Pediatric ICU Neonatal ICU Cardiac ICU Medical and surgical acute care areas
Shea et al, 2013 [40]	Children	36	6,872	Pediatric ICU Cardiac ICU
Stockwell et al, 2013 [25]	Children, exclusion of emergency departments and ambulatory clinics	36 for hospital 1 and 51 for hospital 2	Not stated	General internal medicine Surgical care Psychiatric unit Neonatal Cardiac ICU Pediatric ICU

**Table 2.** Data sources, triggers, and reviewers of included studies.

Study	Data source	Triggers	Reviewer(s)
<b>Diagnostic test accuracy studies</b>			
Gerdes and Hardahl, 2013 [44]	Unstructured and semistructured narrative texts in EHRs <sup>a</sup>	“Models,” not defined, identifying the most common triggers and/or AEs <sup>b</sup>	Not stated
O’Leary et al, 2013 [41]	Enterprise Data Warehouse: EHRs or CPOES <sup>c</sup> ; hospital and physician billing systems; incident reporting system; and admission or discharge or transfer with nightly updates from activities occurring in the preceding 24 h	Locally developed based on screening criteria from the Harvard Medical Practice Study and the IHI <sup>d</sup> GTT <sup>e</sup>	Experienced hospitalists and physician-researcher (prior experience with the research method)
<b>Prevalence studies</b>			
Call et al, 2014 [36]	EHR: laboratory, pharmacy, electronic medication administration record, CPOE, and documentation functions	Wide use in similar population and high likelihood to detect adverse drug events	Pharmacist and physician
Dickermann et al, 2011 [37]	EHRs	Increasing use in hospitals’ protocols	CA <sup>f</sup> trained
Lim et al, 2016 [42]	EHR supports all inpatient and ambulatory care clinical and documentation activities	Review of literature and detectable in EHRs with reasonable PPV <sup>g</sup>	Pharmacists, medication safety pharmacist, and physician
Moore et al, 2009 [38]	CPOE with decision support, EHR, clinical event monitors	Most common inpatient adverse drug events	Study investigators
Muething et al, 2010 [39]	Clinical information system: computerized clinical order entry, clinical documentation, electronic medication administration record, data storage repository, and advanced clinical decision support	AEs steering committee	Endocrinologist, anesthesiologist, and frontline staff
Nwulu et al, 2013 [45]	Locally developed electronic health and prescription computer system (laboratory results and prescribing, except some chemotherapy regimens) has built-in checks to identify potential prescribing errors (flagged through warnings and alerts)	Test the usefulness of two medication module triggers from the GTT proposed by IHI	Not stated
Patregnani et al, 2015 [43]	EHRs	Clinical evidences	CA trained in the AE trigger process
Shea et al, 2013 [40]	EHRs and Laboratory Information System	Clinical evidences and risks of deaths	CA trained in the AE trigger process
Stockwell et al, 2013 [25]	EHRs	Multidisciplinary review process using several review criteria	CA

<sup>a</sup>EHRs: electronic health records.<sup>b</sup>AE: adverse event.<sup>c</sup>CPOES: computerized provider order entry system.<sup>d</sup>IHI: Institute for Healthcare Improvement.<sup>e</sup>GTT: Global Trigger Tool.<sup>f</sup>CA: clinical analyst.<sup>g</sup>PPV: positive predictive value.

**Table 3.** Overview of the automated trigger tool methodology.

Study	Description of the method
<b>Diagnostic test accuracy studies</b>	
Gerdes and Hardahl, 2013 [44]	(1) Extraction and preparation of all texts from the EHRs <sup>a</sup> ; (2) Use of SAS Text Miner and the SAS Enterprise Content Categorization software to build query models (natural language processing algorithms)
O’Leary et al, 2013 [41]	(1) Leveraging of various information systems in the EDW <sup>b</sup> ; (2) Write Structured Query Language queries to mimic work of a reviewer to identify potential AEs <sup>c</sup> based on trigger tool; (3) Two reviewers review the positive EDW screens; (4) Another reviewer reviews narrative summaries and determines presence of AEs
<b>Prevalence studies</b>	
Call et al, 2014 [36]	(1) Software program conducts an extensive search of patient records for any type of order containing specific medications and laboratory values; (2) Information generated into a report with patient-specific information; (3) Review by two reviewers
Dickermann et al, 2011 [37]	(1) Trigger reports automatically generated on a daily basis from the EHR by querying the Sunquest Laboratory Information System for laboratory results; (2) Reviewer examined every trigger by reading the EHRs and interviewing care providers
Lim et al, 2016 [42]	(1) Administration of a trigger drug to a patient automatically sent an electronic trigger-detection message to two reviewers; (2) Trigger-detection messages were evaluated immediately after or during the day by both reviewers (consensus if disagreement); (3) Event reviewed by a medication safety pharmacist and then by a physician for validation.
Moore et al, 2009 [38]	(1) The laboratory results and administered medications of each adult hospital patient were continuously monitored by the computerized trigger alert system; (2) If any of the conditions defined was satisfied (trigger algorithm), an alert was triggered, and data were collected by study investigators on the patient for a period of 72 hours after the initial trigger firing to determine whether an adverse drug event had occurred.
Muething et al, 2010 [39]	(1) Combination of trigger tool approach with the clinical information system; (2) Every evening, automatic detection of triggers are sent to the project manager (detection of event within 24 h); (3) Summary of the incident automatically generated and sent to the appropriate staff on the unit involved
Nwulu et al, 2013 [45]	(1) The triggers identified electronically were linked to the electronic prescription records; (2) Two or more positive triggers generated for the same patient, within a 24- or 72-hour interval (trigger-dependent) were treated as one trigger; (3) The paper-based case notes were reviewed to identify any documentation of interest
Patrengani et al, 2015 [43]	(1) Generation of a trigger report by querying the Laboratory Information System (2) Reviewer investigated the event by reading the patient’s EHRs and interviewing the clinical care team
Shea et al, 2013 [40]	(1) Generation of a trigger report by querying the Laboratory Information System (2) Reviewer investigated the event by reading the patient’s EHRs and interviewing the clinical care team
Stockwell et al, 2013 [25]	(1) Automated trigger reports are generated from hospital information systems on a nightly basis; (2) Each trigger report is examined by a reviewer and interviews conducted with care providers.

<sup>a</sup>EHRs: electronic health records.<sup>b</sup>EDW: Enterprise Data Warehouse.<sup>c</sup>AEs: adverse events.

Two studies compared results from automated trigger tool methodology with those obtained via the manual trigger tool method [41,44]. Six studies compared results with voluntary incident reports but did not cross-verify their results with those obtained via manually operated trigger-based tools; therefore, we considered these prevalence studies [25,36,37,39,43,45]. The remaining 3 studies did not compare their results with those obtained via any other method [38,40,42].

## Quality Assessment

Figure 2 shows detailed results of the two quality assessments.

### Diagnostic Test Accuracy Studies

In one of the 2 studies, the assessment of all domains was hampered by poor reporting, and we were unable to judge the risk of bias and concerns regarding applicability [44]. In the other, we judged the concern regarding applicability of “patient

selection” as low [41]. We judged a high risk of bias in the “flow and timing” domain and low bias risk and applicability concerns for the “index test” and “reference standard” domains [41].

### Prevalence Studies

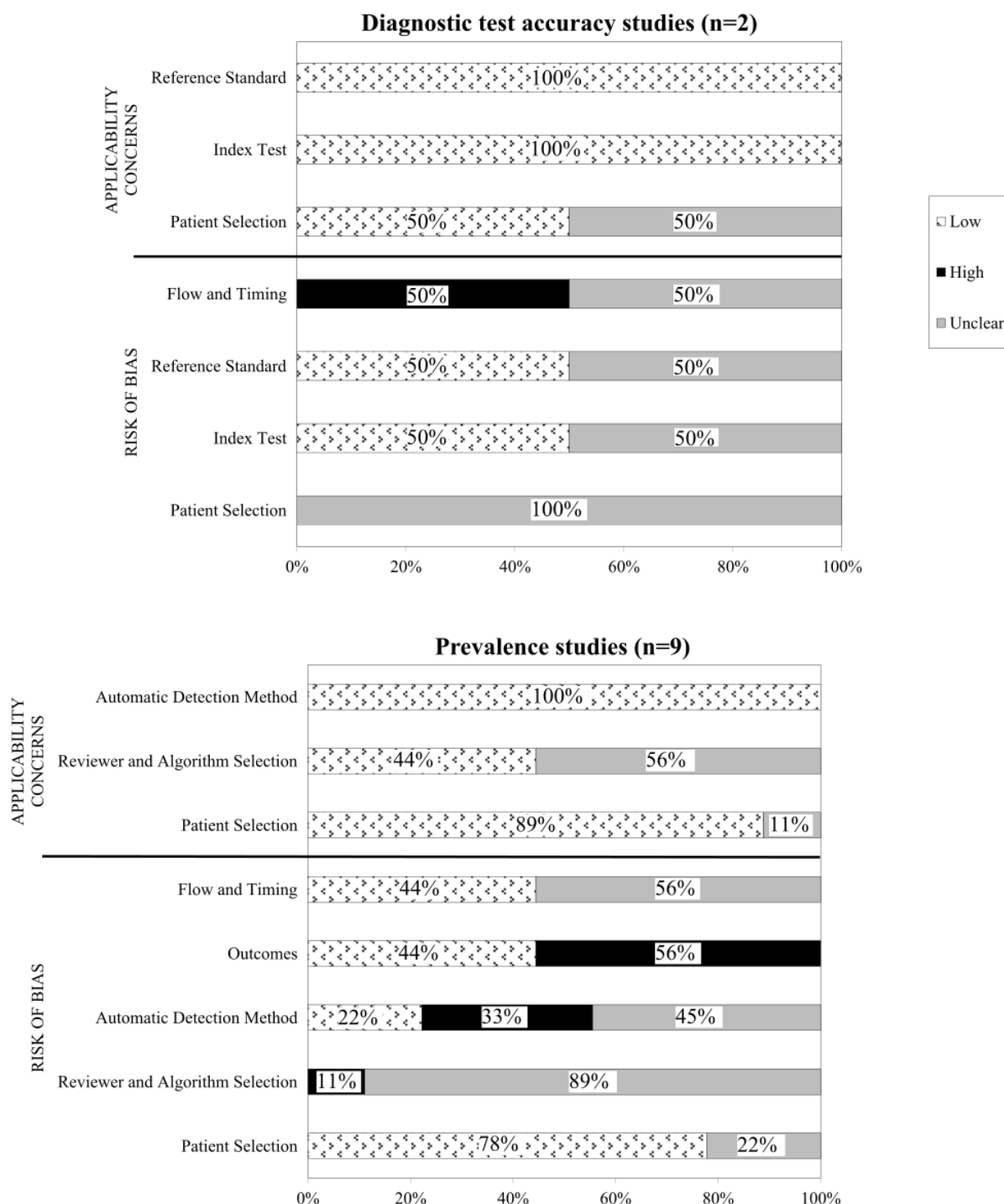
A total of 9 studies were selected as prevalence studies via our self-developed quality assessment tool (Multimedia Appendix 3). For “patient selection,” the bias risk was deemed low in seven studies [25,36-38,40,42,43] and applicability concerns were also low in eight of them [25,36-40,42,43]. We had no concerns regarding applicability of the automatic detection methods. As Figure 2 shows, for the other domain, we judged low bias risk and concerns regarding applicability in a minority of studies. In 5 studies, we judged a high risk of bias in the “outcomes” domain, mainly because their AE definitions did not reference those of the IHI or Food and Drug Administration, and they lacked clearly stated prevalence outcomes

[37,38,42,43,45]. We judged high bias risks in 3 studies regarding the “automated detection method” domain [25,37,45]. In another, we judged the bias risk as high in the “reviewer or algorithm selection” domain [45] because the triggers were not consistently used or developed, and the reviewer lacked the required profile (eg, experience and training).

### Estimates of Diagnostic Test Accuracy, Prevalence, and Reliability

Estimates of DTA can be found in Table 4, with additional information concerning prevalence, preventability, and false negative rates for AE categories. No information concerning 2x2 tables were available for all the triggers; thus, the decision was made to use the AE categories.

**Figure 2.** Risk of bias and concerns regarding applicability assessments for diagnostic test accuracy studies (upper panel) and prevalence studies (lower panel). Judgments are expressed as “low,” “high,” or “unclear” risk or concern for each of the domains (ie, “patient selection,” “index test”). The percentages refer to the percentage of studies meeting the judgment low, high, or unclear risk of bias or concerns regarding applicability in each of the domains. Quality Assessment tool for Diagnostic Accuracy Studies-2 (QUADAS-2) was used for the two diagnostic test accuracy studies and an in-house developed tool was used to assess the 9 prevalence studies.





**Table 4.** The table displays the estimates of diagnostic test accuracy in 2 studies comparing automated trigger-based tools with a manual trigger-based tool as reference standard.

Study	Type of adverse events	2x2 table for adverse events (True positive / false positive / false negative / true negative)	Prevalence <sup>a</sup> , % (95% CI)	Positive predictive value <sup>b</sup> , % (95% CI)	False negative rate <sup>c</sup> (%)
Gerdes and Hardahl, 2013 [44]	Pressure ulcer	28 / 22 / 12 / 436	5.6 (3.6-7.6)	56 (42.2-69.8)	30
O'Leary et al, 2013 [41]	Adverse drug event	24 / 22 / 20 / N/A <sup>d</sup>	9.6 (5.9-13.3)	52.2 (37.7-66.6)	45.5
	Hospital acquired infection	7 / 11 / 4 / N/A	2.8 (0.8-4.9)	38.9 (16.4-61.4)	36.4
	Operative or procedural injury	5 / 4 / 4 / N/A	2 (0.3-3.7)	55.6 (23.1-88)	44.4
	Manifestation of poor glycemic control	3 / 2 / 5 / N/A	1.2 (-0.2 to 2.6)	60 (17.1-102.9)	62.5
	Pressure ulcer	0 / 8 / 2 / N/A	0 (0-0)	0 (0-0)	100
	Venous thromboembolism	5 / 1 / 0 / N/A	2 (0.3-3.7)	83.3 (53.5-113.2)	0
	Acute renal failure	2 / 1 / 0 / N/A	0.8 (-0.3 to 1.9)	66.7 (13.3-120)	0
	Delirium	0 / 0 / 0 / N/A	0 (0-0)	0 (0-0)	0
	Fall	0 / 0 / 0 / N/A	0 (0-0)	0 (0-0)	0
	Other	0 / 2 / 5 / N/A	0 (0-0)	0 (0-0)	100

<sup>a</sup>Prevalence is calculated by true positive/total number of patients.

<sup>b</sup>Calculated as triggers corresponding to an adverse event out of all triggers=true positive/(true positive+false positive).

<sup>c</sup>Calculated as false negative/(false negative+true positive).

<sup>d</sup>N/A: not applicable.

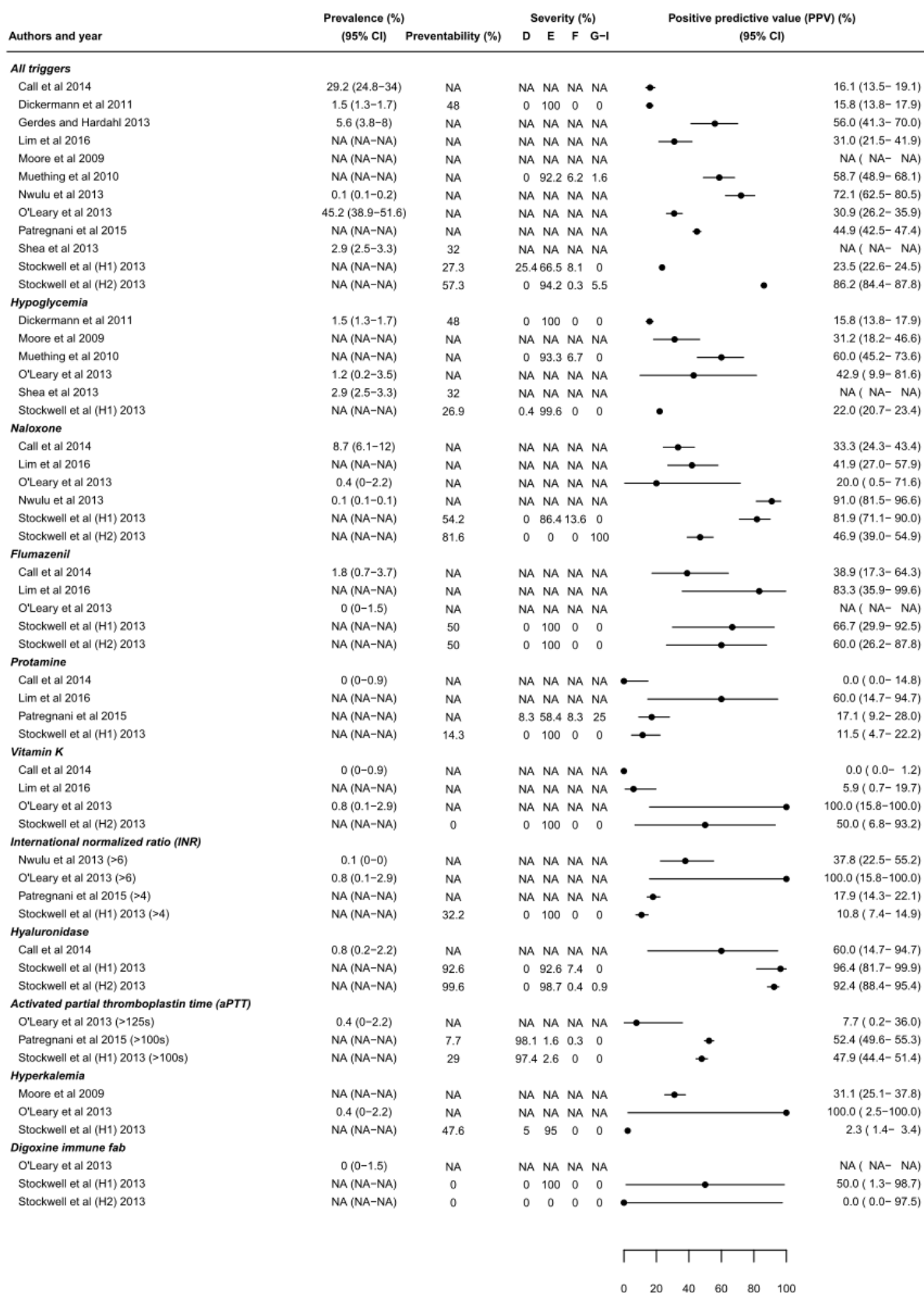
Across all the 11 studies, AE prevalence ranged from 0% to 17.9%, with a median of 0.8%. The PPV of all triggers to detect AEs ranged from 0% to 100% across studies, with a median of 40%. Some triggers are used in different studies and have different PPV values: (1) in 6 studies, hypoglycemia [25,37-41] had a PPV ranging from 15.8% to 60%; (2) in 5 studies, naloxone [25,36,41,42,45] had a PPV ranging from 20% to 91%; (3) in 4 studies, flumazenil [25,36,41,42] had a PPV ranging from 38.9% to 83.3%; and (4) in 4 studies, protamine [25,36,42,43] had a PPV ranging from 0% to 60%. We were unable to determine the AE prevalence, PPV, preventability, and severity in 40.4%, 10.5%, 71.1%, and 68.4% of the studies, respectively. These studies did not report the overall number of records analyzed, triggers, or AEs; or the studies did not conducted the analysis concerned. Detailed results for each trigger with prevalence, preventability, severity, and PPV are presented in Figures 3 and 4. Only 1 study supplied information

on interrater reliability, reporting a kappa value of .52 to .78 [41].

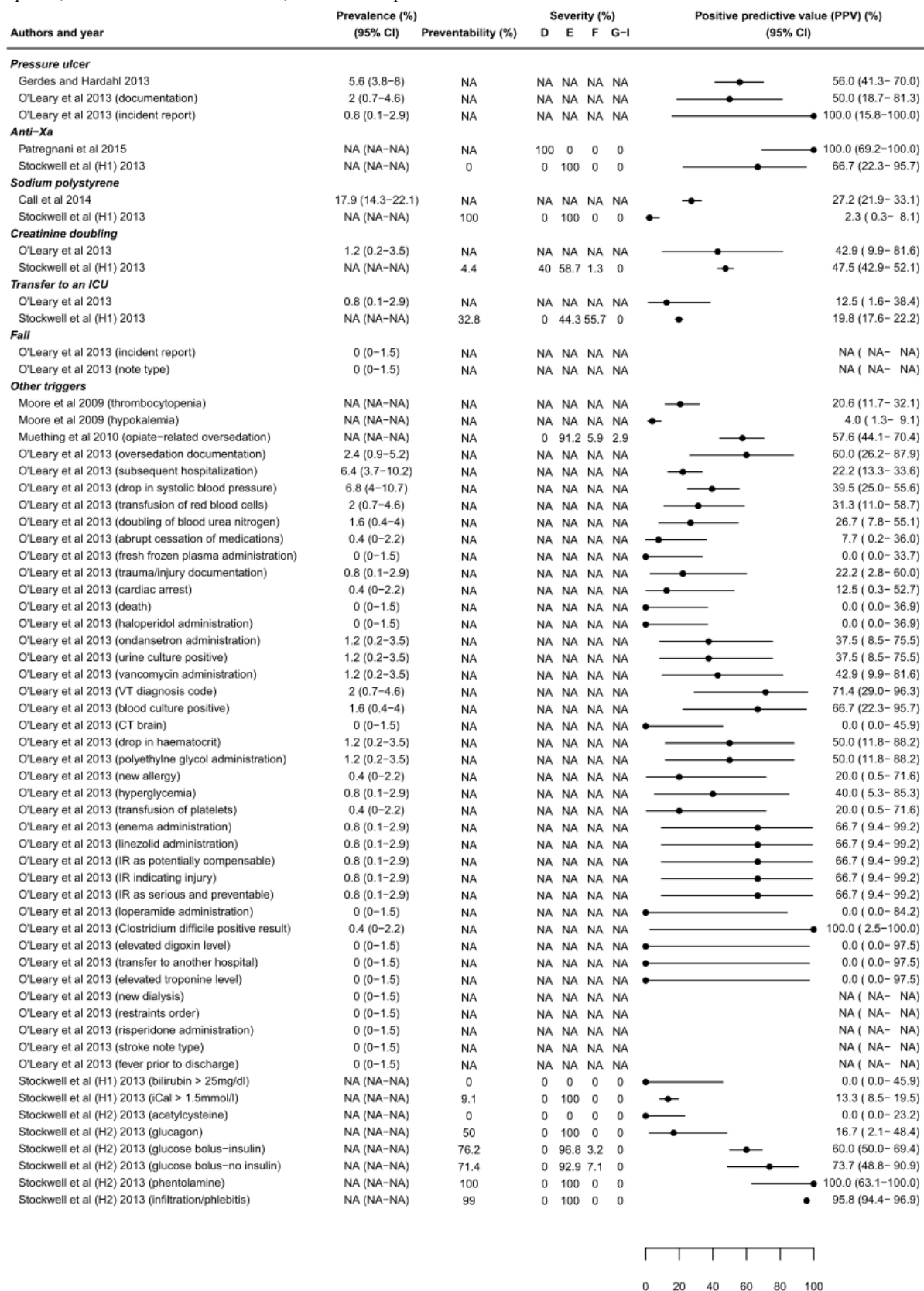
### Challenges and Author Proposed Solutions

The challenges reported by authors concerned the relative large number of false alarms, the dependence of PPV on AE prevalence, and incomplete patient records leading to missed events. To reduce the fraction of false alarms, several authors suggested to use a threshold value for the triggers used based on patient characteristics [36,38,40,43]. For example, specific disease states of patients must have triggers with different threshold as the consequences might be stronger because of their disease. Authors from one study suggested to measure sensitivity and specificity instead of PPV, as latter is known to be largely affected by AE prevalence [38]. Another author group suggested to either improve completeness of patient information documentation or to combine different data sources to increase the chance to detect AEs as each data source contains different information type [42].

**Figure 3.** Prevalence, preventability, severity, and positive predictive value (PPV) for all the 11 studies. The figure begins with the results of all the triggers or adverse events (AEs) combined, then for each group of trigger order from the most studied to the least studied (part 1). Severity levels based on the National Coordinating Council for Medication Error Reporting and Prevention: D=an error that reached the patient and required monitoring or intervention to confirm that it resulted in no harm to the patient; E=temporary harm to the patient and required intervention; F=temporary harm to the patient and required initial or prolonged hospitalization; G=permanent patient harm; H=intervention required to sustain life; and I=patient death. H1: hospital 1; H2: hospital 2.



**Figure 4.** Prevalence, preventability, severity, and positive predictive value (PPV) for all the 11 studies. The figure begins with the results of all the triggers or adverse events (AEs) combined, then for each group of trigger order from the most studied to the least studied (part 2). Severity levels based on the National Coordinating Council for Medication Error Reporting and Prevention: D=an error that reached the patient and required monitoring or intervention to confirm that it resulted in no harm to the patient; E=temporary harm to the patient and required intervention; F=temporary harm to the patient and required initial or prolonged hospitalization; G=permanent patient harm; H=intervention required to sustain life; I=patient death. H1: hospital 1; H2: hospital 2; VT: venous thromboembolism; IR: incident report.



## Discussion

### Aim of This Review and Principal Findings

The goal of this systematic review was to synthesize the evidence concerning the development of a semi or fully automated method of AE detection in EHRs using trigger tools. The results show a broad variation in applied methods, selection of triggers, and estimates of AE prevalence and trigger-based PPVs. Insufficient reporting precluded full appreciation of the risk of bias and concerns regarding applicability.

Our findings are in line with another systematic review focusing on manual GTT [10]. Several interstudy differences can be hypothesized to explain the heterogeneity in the observed study estimates of AE prevalence and PPVs. These include the time frame (range: 1-51 months); the sample size (range: 250-54,244 records); the data sources and EHR system types; the automated approach; the review process; but also the differences in defining AEs, its severity, and preventability.

In addition, the choice and definition of triggers (eg, INR  $\geq 6$  [41,45] and INR  $> 4$  [25,43]) affect overall and per-trigger PPV. In some studies, only triggers identifying *unique* AEs are used for analysis, leading to varying results for similar triggers. Furthermore, the PPV is deemed to be strongly affected by the study's AE prevalence. We could not test for this because of the reviewed studies' heterogeneous definitions of prevalence. These included AEs per 100 patients admitted, AEs per 1000 patient days, or AEs per month. The missing information concerning the total number of patients included (36.4% of the 11 studies) hindered computation of prevalence estimates for these studies. The PPV of the triggers ranged from 0% to 100%. Even for the same trigger, high variability was observed.

Similar parallels can be observed regarding AEs' severity. Even when naloxone was used as a trigger in two hospitals in the same study [25], although one hospital had a severity level of 96.4%, falling into category "temporary harm to the patient and required intervention," the other had a severity level 100%, falling into the categories "permanent patient harm or patient death." Several included studies suggested improvement of PPV values by limiting triggers to specific patient subpopulations, or by modifying thresholds based on age or other patient characteristics [36,40]. Moore et al [38] suggested using sensitivity and specificity instead of PPV, as the former is less affected by changes in AE incidence. The studies did not address difficulties or opportunities regarding the implementation of trigger tool-based methods, but we do not exclude that such information can be found in more qualitative types of research that we did not collect for this review.

The decision of several of the reviewed studies' authors to employ semiautomatic approaches or to focus mainly on ADEs reflects the difficulty of detecting AEs fully automatically. The extracted high proportion of false alarms in various studies is likely explained in part by these difficulties. Although several studies showed that automatic tools can detect up to 69% of ADEs found manually, automatic detection of AEs overall remains problematic [47-50]. In our review, the sensitivity to detect AEs ranged from 0% to 100%. General AE detection

requires more advanced technology, such as NLP, which can read and process free-text narrative [51,52], addressing complex issues such as negation and lexical variation of terminology. Previous research has suggested that automated AE detection methods were superior to manual tools [26,53,54]. Automated AE detection methods have the potential to screen large numbers of patients to save valuable time, which would not be possible by doing manually by human reviewers with the same accuracy. Yet, timely intervention is an important factor to avoid complications and patient harm when an AE is detected. Even if PPVs are not high for all AEs, automated AE detection methods provide an excellent alternative to the manual approach by saving time and resources [14]. As shown by the systematic review of Wang et al, the use of NLP with EHRs is still at its infancy, and closer collaboration of NLP experts and clinicians is missing [46]. Nevertheless, automated AE detection methods are a promising approach for patient safety improvement.

Looking deeper into the individual studies via our quality assessment tools, we realized that, for most, their methodology, their results, or both were reported in ways that were unclear, inconsistent, or incomplete, which challenged our risk of bias and applicability assessments. Several studies failed to report the number of records screened or the type of patients sampled. These elements, however, are crucial for interpretation of the various estimates and also for its reproducibility. In other studies, the number of triggers or AEs remained unclear. AE severity and preventability were not always reported, and the variation in use of definitions for AEs, severity, and preventability further hampered interpretation of estimates across studies. Interestingly, the majority of studies lacked to report PPV CIs, which is essential for the swift interpretation of the estimate's precision.

Although not part of our quality assessment tools, we detected risk of selective outcome reprint in some studies. In these, severity and/or preventability assessments are reported in the Methods sections but not addressed in the Results section. Only one study checked for the presence of triggers at admission [42]. Such triggers or AEs should typically be excluded from the evaluation, as it cannot be targeted with interventions aiming at improving quality of care. Furthermore, although a reviewer's expertise plays an important role in the detection of AEs, information concerning their professional background, experience, or training was mostly absent. Overall, there is substantial room for improvement of the quality of reporting.

### Limitations

Our decision to limit the inclusion criteria to studies concerning semi or fully automated trigger tool-like methodologies disqualified many studies, including those employing recommendations from the Harvard Medical Practice Study [28,55,56], machine learning [57], early warning systems [58-61], or other methods [62,63]. However, it allowed us to show that even within a narrow set of trigger-based tools, methods and outcomes varied considerably. The decision to exclude studies involving outpatients [62,64,65] or mixes of inpatients and outpatients [66] further decreased the number of eligible studies but increased the comparability of the patient population evaluated. Nevertheless, the overall low number of



eligible studies precluded statistical evaluation of the impact of sources of variation and bias. The evaluation of diagnostic test accuracy is generally hampered by the absence of a widely accepted reference standard.

## Conclusions

This systematic review provides an overview about the application and outcomes of (semi) automatic trigger-based AE detection tools. We observed but could not formally explain the high degree of interstudy variation in reported estimates of prevalence and PPV, even in cases where similar triggers were tested. Although the AHRQ recently released common formats for event reporting [67], which supports the implementation of AE detection in the EHR, standards for the reporting of AE detection studies using trigger tools are lacking, yet urgently

needed to overcome the methodological heterogeneity in future studies. We need better standards for reporting in this field of research to increase reproducibility, interpretation, and avoidance of research waste. A more standardized use of definitions of the types, severity, and preventability of AEs is desirable. We therefore call for a set of recommendations for the conduct and reporting of future studies and in the meantime, suggest authors, peer reviewers, and editors to pay special attention to complete reporting of study population, AE and trigger definitions, experience, training, and background of reviewers; methods employed to check for triggers and/or AEs at patient admission; and complete reporting of outcome data (numbers of triggers, nominators and denominators of the prevalence, and PPV).

## Acknowledgments

The authors would like to thank Barbara Schärer for her assistance in data extraction and quality assessment as part of her Master studies.

## Authors' Contributions

MS developed the idea for the study. MS, SNM, and AWSR contributed to the concept, design, analysis, and interpretation of data. SNM contributed to the drafting of the manuscript. AWSR contributed to the analysis and interpretation of the data. DA, RS, HUR, MMJ, and AWSR contributed to the screening and extraction phases, the interpretation of data, and the critical revision of the manuscript. All authors approved the final version.

## Conflicts of Interest

None declared.

## Multimedia Appendix 1

Hausner et al's approach.

[PDF File (Adobe PDF File), 16KB - [jmir\\_v20i5e198\\_app1.pdf](#)]

## Multimedia Appendix 2

Quality Assessment tool for Diagnostic Accuracy Studies–2 (QUADAS–2) instrument.

[PDF File (Adobe PDF File), 42KB - [jmir\\_v20i5e198\\_app2.pdf](#)]

## Multimedia Appendix 3

Trigger tool Quality Assessment Tool.

[PDF File (Adobe PDF File), 42KB - [jmir\\_v20i5e198\\_app3.pdf](#)]

## References

1. Woloshynowych M, Neale G, Vincent C. Case record review of adverse events: a new approach. *Qual Saf Health Care* 2003 Dec;12(6):411-415 [FREE Full text] [Medline: [14645755](#)]
2. Vincent C, Neale G, Woloshynowych M. Adverse events in British hospitals: preliminary retrospective record review. *Br Med J* 2001 Mar 3;322(7285):517-519 [FREE Full text] [Medline: [11230064](#)]
3. Institute of Medicine. In: Kohn LT, Corrigan JM, Donaldson MS, editors. *To Err is Human: Building a Safer Health System*. Washington, DC: The National Academies Press; 2000.
4. Classen D. Medication safety: moving from illusion to reality. *J Am Med Assoc* 2003 Mar 05;289(9):1154-1156. [Medline: [12622587](#)]
5. Adler L, Yi D, Li M, McBroom B, Hauck L, Sammer C, et al. Impact of Inpatient Harms on Hospital Finances and Patient Clinical Outcomes. *J Patient Saf* 2015 Mar 23:- (Forthcoming)(forthcoming). [doi: [10.1097/PTS.000000000000171](#)] [Medline: [25803176](#)]



6. Baker GR, Norton PG, Flintoft V, Blais R, Brown A, Cox J, et al. The Canadian Adverse Events Study: the incidence of adverse events among hospital patients in Canada. *CMAJ* 2004 May 25;170(11):1678-1686 [[FREE Full text](#)] [Medline: [15159366](#)]
7. Brennan TA, Leape LL, Laird NM, Hebert L, Localio AR, Lawthers AG, Harvard Medical Practice Study I. Incidence of adverse events and negligence in hospitalized patients: results of the Harvard Medical Practice Study I. 1991. *Qual Saf Health Care* 2004 Apr;13(2):145-51; discussion 151 [[FREE Full text](#)] [Medline: [15069223](#)]
8. Leape LL, Brennan TA, Laird N, Lawthers AG, Localio AR, Barnes BA, et al. The nature of adverse events in hospitalized patients. Results of the Harvard Medical Practice Study II. *N Engl J Med* 1991 Feb 7;324(6):377-384. [doi: [10.1056/NEJM199102073240605](#)] [Medline: [1824793](#)]
9. Thomas EJ, Studdert DM, Burstin HR, Orav EJ, Zeena T, Williams EJ, et al. Incidence and types of adverse events and negligent care in Utah and Colorado. *Med Care* 2000 Mar;38(3):261-271. [Medline: [10718351](#)]
10. Hibbert PD, Molloy CJ, Hooper TD, Wiles LK, Runciman WB, Lachman P, et al. The application of the Global Trigger Tool: a systematic review. *Int J Qual Health Care* 2016 Dec 01;28(6):640-649. [doi: [10.1093/intqhc/mzw115](#)] [Medline: [27664822](#)]
11. Rosen AK, Mull HJ. Identifying adverse events after outpatient surgery: improving measurement of patient safety. *BMJ Qual Saf* 2016 Jan;25(1):3-5. [doi: [10.1136/bmjqs-2015-004752](#)] [Medline: [26453634](#)]
12. Gandhi TK, Seger DL, Bates DW. Identifying drug safety issues: from research to practice. *Int J Qual Health Care* 2000 Feb;12(1):69-76. [Medline: [10733086](#)]
13. Classen DC, Resar R, Griffin F, Federico F, Frankel T, Kimmel N, et al. 'Global trigger tool' shows that adverse events in hospitals may be ten times greater than previously measured. *Health Aff (Millwood)* 2011 Apr;30(4):581-589 [[FREE Full text](#)] [doi: [10.1377/hlthaff.2011.0190](#)] [Medline: [21471476](#)]
14. Murff HJ, Patel VL, Hripcsak G, Bates DW. Detecting adverse events for patient safety research: a review of current methodologies. *J Biomed Inform* 2003;36(1-2):131-143 [[FREE Full text](#)] [Medline: [14552854](#)]
15. Ausserhofer D, Zander B, Busse R, Schubert M, De Geest S, Rafferty AM, RN4CAST consortium. Prevalence, patterns and predictors of nursing care left undone in European hospitals: results from the multicountry cross-sectional RN4CAST study. *BMJ Qual Saf* 2014 Feb;23(2):126-135. [doi: [10.1136/bmjqs-2013-002318](#)] [Medline: [24214796](#)]
16. Berry LL, Segal R, Sherrin TP, Fudge KA. Sensitivity and specificity of three methods of detecting adverse drug reactions. *Am J Hosp Pharm* 1988 Jul;45(7):1534-1539. [Medline: [3046347](#)]
17. Rogers AS, Israel E, Smith CR, Levine D, McBean AM, Valente C, et al. Physician knowledge, attitudes, and behavior related to reporting adverse drug events. *Arch Intern Med* 1988 Jul;148(7):1596-1600. [Medline: [3382304](#)]
18. Rozich JD, Haraden CR, Resar RK. Adverse drug event trigger tool: a practical methodology for measuring medication related harm. *Qual Saf Health Care* 2003 Jun;12(3):194-200 [[FREE Full text](#)] [Medline: [12792009](#)]
19. Resar RK, Rozich JD, Simmonds T, Haraden CR. A trigger tool to identify adverse events in the intensive care unit. *Jt Comm J Qual Patient Saf* 2006 Oct;32(10):585-590. [Medline: [17066996](#)]
20. Cullen DJ, Bates DW, Small SD, Cooper JB, Nemeskal AR, Leape LL. The incident reporting system does not detect adverse drug events: a problem for quality improvement. *Jt Comm J Qual Improv* 1995 Oct;21(10):541-548. [Medline: [8556111](#)]
21. West AN, Weeks WB, Bagian JP. Rare adverse medical events in VA inpatient care: reliability limits to using patient safety indicators as performance measures. *Health Serv Res* 2008 Feb;43(1 Pt 1):249-266 [[FREE Full text](#)] [doi: [10.1111/j.1475-6773.2007.00760.x](#)] [Medline: [18211528](#)]
22. Griffin FA, Resar RK. IHI. 2009. IHI Global Trigger Tool for Measuring Adverse Events URL: <http://www.ihl.org/resources/Pages/Tools/IHIGlobalTriggerToolforMeasuringAEs.aspx> [accessed 2018-05-04] [[WebCite Cache ID 6z9v8yzwv](#)]
23. Adler L, Denham CR, McKeever M, Purinton R, Guilloteau F, Moorhead JD, et al. Global trigger tool: implementation basics. *J Patient Saf* 2008;4(4):245-249. [doi: [10.1097/pts.0b013e31818e8a87](#)]
24. Doupi P. Using EHR data for monitoring and promoting patient safety: reviewing the evidence on trigger tools. *Stud Health Technol Inform* 2012;180:786-790. [Medline: [22874299](#)]
25. Stockwell DC, Kirkendall E, Muething SE, Kloppenborg E, Vinodrao H, Jacobs BR. Automated adverse event detection collaborative: electronic adverse event identification, classification, and corrective actions across academic pediatric institutions. *J Patient Saf* 2013 Dec;9(4):203-210. [doi: [10.1097/PTS.0000000000000055](#)] [Medline: [24257063](#)]
26. Li Q, Melton K, Lingren T, Kirkendall ES, Hall E, Zhai H, et al. Phenotyping for patient safety: algorithm development for electronic health record based automated adverse event and medical error detection in neonatal intensive care. *J Am Med Inform Assoc* 2014;21(5):776-784 [[FREE Full text](#)] [doi: [10.1136/amiajnl-2013-001914](#)] [Medline: [24401171](#)]
27. Menendez ME, Janssen SJ, Ring D. Electronic health record-based triggers to detect adverse events after outpatient orthopaedic surgery. *BMJ Qual Saf* 2016 Jan;25(1):25-30. [doi: [10.1136/bmjqs-2015-004332](#)] [Medline: [26359272](#)]
28. Murff HJ, Forster AJ, Peterson JF, Fiskio JM, Heiman HL, Bates DW. Electronically screening discharge summaries for adverse medical events. *J Am Med Inform Assoc* 2003;10(4):339-350 [[FREE Full text](#)] [doi: [10.1197/jamia.M1201](#)] [Medline: [12668691](#)]

29. Govindan M, Van Citters AD, Nelson EC, Kelly-Cummings J, Suresh G. Automated detection of harm in healthcare with information technology: a systematic review. *Qual Saf Health Care* 2010 Oct;19(5):e11. [doi: [10.1136/qshc.2009.033027](https://doi.org/10.1136/qshc.2009.033027)] [Medline: [20671081](https://pubmed.ncbi.nlm.nih.gov/20671081/)]
30. Reitsma JB, Rutjes AW, Whiting P, Vlasov VV, Leeflang MMG, Deeks JJ. Chapter 9: Assessing methodological quality. In: *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy*. London, England: The Cochrane Collaboration; 2009.
31. Higgins JP, Green S, editors. *Cochrane Handbook for Systematic Reviews of Interventions*. West Sussex, England: John Wiley & Sons Ltd; 2011.
32. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med* 2009 Jul 21;6(7):e1000097 [FREE Full text] [doi: [10.1371/journal.pmed.1000097](https://doi.org/10.1371/journal.pmed.1000097)] [Medline: [19621072](https://pubmed.ncbi.nlm.nih.gov/19621072/)]
33. Hausner E, Waffenschmidt S, Kaiser T, Simon M. Routine development of objectively derived search strategies. *Syst Rev* 2012 Feb 29;1:19 [FREE Full text] [doi: [10.1186/2046-4053-1-19](https://doi.org/10.1186/2046-4053-1-19)] [Medline: [22587829](https://pubmed.ncbi.nlm.nih.gov/22587829/)]
34. Hausner E, Guddat C, Hermanns T, Lampert U, Waffenschmidt S. Development of search strategies for systematic reviews: validation showed the noninferiority of the objective approach. *J Clin Epidemiol* 2015 Feb;68(2):191-199 [FREE Full text] [doi: [10.1016/j.jclinepi.2014.09.016](https://doi.org/10.1016/j.jclinepi.2014.09.016)] [Medline: [25464826](https://pubmed.ncbi.nlm.nih.gov/25464826/)]
35. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011 Oct 18;155(8):529-536. [doi: [10.7326/0003-4819-155-8-201110180-00009](https://doi.org/10.7326/0003-4819-155-8-201110180-00009)] [Medline: [22007046](https://pubmed.ncbi.nlm.nih.gov/22007046/)]
36. Call RJ, Burlison JD, Robertson JJ, Scott JR, Baker DK, Rossi MG, et al. Adverse drug event detection in pediatric oncology and hematology patients: using medication triggers to identify patient harm in a specialized pediatric patient population. *J Pediatr* 2014 Sep;165(3):447-52.e4 [FREE Full text] [doi: [10.1016/j.jpeds.2014.03.033](https://doi.org/10.1016/j.jpeds.2014.03.033)] [Medline: [24768254](https://pubmed.ncbi.nlm.nih.gov/24768254/)]
37. Dickerman MJ, Jacobs BR, Vinodrao H, Stockwell DC. Recognizing hypoglycemia in children through automated adverse-event detection. *Pediatrics* 2011 Apr;127(4):e1035-e1041. [doi: [10.1542/peds.2009-3432](https://doi.org/10.1542/peds.2009-3432)] [Medline: [21402631](https://pubmed.ncbi.nlm.nih.gov/21402631/)]
38. Moore C, Li J, Hung C, Downs J, Nebeker JR. Predictive value of alert triggers for identification of developing adverse drug events. *J Patient Saf* 2009 Dec;5(4):223-228. [doi: [10.1097/PTS.0b013e3181bc05e5](https://doi.org/10.1097/PTS.0b013e3181bc05e5)] [Medline: [22130215](https://pubmed.ncbi.nlm.nih.gov/22130215/)]
39. Muething SE, Conway PH, Kloppenborg E, Lesko A, Schoettker PJ, Seid M, et al. Identifying causes of adverse events detected by an automated trigger tool through in-depth analysis. *Qual Saf Health Care* 2010 Oct;19(5):435-439. [doi: [10.1136/qshc.2008.031393](https://doi.org/10.1136/qshc.2008.031393)] [Medline: [20798069](https://pubmed.ncbi.nlm.nih.gov/20798069/)]
40. Shea PL, Spaeder MC, Sayal P, Jacobs BR, Stockwell DC. Hypoglycemia adverse events in PICUs and cardiac ICUs: differentiating preventable and nonpreventable events\*. *Pediatr Crit Care Med* 2013 Oct;14(8):741-746. [doi: [10.1097/PCC.0b013e3182975f0f](https://doi.org/10.1097/PCC.0b013e3182975f0f)] [Medline: [23863820](https://pubmed.ncbi.nlm.nih.gov/23863820/)]
41. O'Leary KJ, Devisetty VK, Patel AR, Malkenson D, Sama P, Thompson WK, et al. Comparison of traditional trigger tool to data warehouse based screening for identifying hospital adverse events. *BMJ Qual Saf* 2013 Feb;22(2):130-138. [doi: [10.1136/bmjqs-2012-001102](https://doi.org/10.1136/bmjqs-2012-001102)] [Medline: [23038408](https://pubmed.ncbi.nlm.nih.gov/23038408/)]
42. Lim D, Melucci J, Rizer MK, Prier BE, Weber RJ. Detection of adverse drug events using an electronic trigger tool. *Am J Health Syst Pharm* 2016 Sep 01;73(17 Suppl 4):S112-S120. [doi: [10.2146/ajhp150481](https://doi.org/10.2146/ajhp150481)] [Medline: [27543596](https://pubmed.ncbi.nlm.nih.gov/27543596/)]
43. Patregnani JT, Spaeder MC, Lemon V, Diab Y, Klugman D, Stockwell DC. Monitoring the harm associated with use of anticoagulants in pediatric populations through trigger-based automated adverse-event detection. *Jt Comm J Qual Patient Saf* 2015 Mar;41(3):108-114. [Medline: [25977126](https://pubmed.ncbi.nlm.nih.gov/25977126/)]
44. Gerdes LU, Hardahl C. Text mining electronic health records to identify hospital adverse events. *Stud Health Technol Inform* 2013;192:1145. [Medline: [23920919](https://pubmed.ncbi.nlm.nih.gov/23920919/)]
45. Nwulu U, Nirantharakumar K, Odesanya R, McDowell SE, Coleman JJ. Improvement in the detection of adverse drug events by the use of electronic health and prescription records: an evaluation of two trigger tools. *Eur J Clin Pharmacol* 2013 Feb;69(2):255-259. [doi: [10.1007/s00228-012-1327-1](https://doi.org/10.1007/s00228-012-1327-1)] [Medline: [22706621](https://pubmed.ncbi.nlm.nih.gov/22706621/)]
46. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: a literature review. *J Biomed Inform* 2018 Jan;77:34-49 [FREE Full text] [doi: [10.1016/j.jbi.2017.11.011](https://doi.org/10.1016/j.jbi.2017.11.011)] [Medline: [29162496](https://pubmed.ncbi.nlm.nih.gov/29162496/)]
47. Jha AK, Kuperman GJ, Teich JM, Leape L, Shea B, Rittenberg E, et al. Identifying adverse drug events: development of a computer-based monitor and comparison with chart review and stimulated voluntary report. *J Am Med Inform Assoc* 1998;5(3):305-314 [FREE Full text] [Medline: [9609500](https://pubmed.ncbi.nlm.nih.gov/9609500/)]
48. Bates DW, Leape LL, Cullen DJ, Laird N, Petersen LA, Teich JM, et al. Effect of computerized physician order entry and a team intervention on prevention of serious medication errors. *J Am Med Assoc* 1998 Oct 21;280(15):1311-1316. [Medline: [9794308](https://pubmed.ncbi.nlm.nih.gov/9794308/)]
49. Classen DC, Pestotnik SL, Evans RS, Burke JP. Computerized surveillance of adverse drug events in hospital patients. *J Am Med Assoc* 1991 Nov 27;266(20):2847-2851. [Medline: [1942452](https://pubmed.ncbi.nlm.nih.gov/1942452/)]
50. Raschke RA, Gollhare B, Wunderlich TA, Guidry JR, Leibowitz AI, Peirce JC, et al. A computer alert system to prevent injury from adverse drug events: development and evaluation in a community teaching hospital. *J Am Med Assoc* 1998 Oct 21;280(15):1317-1320. [Medline: [9794309](https://pubmed.ncbi.nlm.nih.gov/9794309/)]

51. Hripcsak G, Friedman C, Alderson PO, DuMouchel W, Johnson SB, Clayton PD. Unlocking clinical data from narrative reports: a study of natural language processing. *Ann Intern Med* 1995 May 01;122(9):681-688. [Medline: [7702231](#)]
52. Friedman C. Towards a comprehensive medical language processing system: methods and issues. *Proc AMIA Annu Fall Symp* 1997:595-599 [FREE Full text] [Medline: [9357695](#)]
53. Murff HJ, FitzHenry F, Matheny ME, Gentry N, Kotter KL, Crimin K, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. *J Am Med Assoc* 2011 Aug 24;306(8):848-855. [doi: [10.1001/jama.2011.1204](#)] [Medline: [21862746](#)]
54. Melton GB, Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. *J Am Med Inform Assoc* 2005;12(4):448-457 [FREE Full text] [doi: [10.1197/jamia.M1794](#)] [Medline: [15802475](#)]
55. Szekendi MK, Sullivan C, Bobb A, Feinglass J, Rooney D, Barnard C, et al. Active surveillance using electronic triggers to detect adverse events in hospitalized patients. *Qual Saf Health Care* 2006 Jun;15(3):184-190 [FREE Full text] [doi: [10.1136/qshc.2005.014589](#)] [Medline: [16751468](#)]
56. Kane-Gill SL, MacLasco AM, Saul MI, Politz ST, Kloet MA, Kim C, et al. Use of text searching for trigger words in medical records to identify adverse drug reactions within an intensive care unit discharge summary. *Appl Clin Inform* 2016 Jul 13;7(3):660-671 [FREE Full text] [doi: [10.4338/ACI-2016-03-RA-0031](#)] [Medline: [27453336](#)]
57. Marella WM, Sparnon E, Finley E. Screening electronic health record-related patient safety reports using machine learning. *J Patient Saf* 2017 Mar;13(1):31-36. [doi: [10.1097/PTS.0000000000000104](#)] [Medline: [24721977](#)]
58. Kavanaugh MJ, So JD, Park PJ, Davis KL. Validation of the intensive care unit early warning dashboard: quality improvement utilizing a retrospective case-control evaluation. *Telemed J E Health* 2017 Feb;23(2):88-95. [doi: [10.1089/tmj.2016.0073](#)] [Medline: [27391204](#)]
59. Le Lagadec MD, Dwyer T. Scoping review: the use of early warning systems for the identification of in-hospital patients at risk of deterioration. *Aust Crit Care* 2017 Jul;30(4):211-218. [doi: [10.1016/j.aucc.2016.10.003](#)] [Medline: [27863876](#)]
60. Paliwoda M, New K, Bogossian F. Neonatal Early Warning Tools for recognising and responding to clinical deterioration in neonates cared for in the maternity setting: a retrospective case-control study. *Int J Nurs Stud* 2016 Sep;61:125-135. [doi: [10.1016/j.ijnurstu.2016.06.006](#)] [Medline: [27348358](#)]
61. Sefton G, Lane S, Killen R, Black S, Lyon M, Sproule C, et al. Accuracy and efficiency of recording paediatric early warning scores (PEWS) using electronic hand-held devices compared with traditional paper-based documentation. *Comput Inform Nurs* 2017 May;35(5):228-236 [FREE Full text] [doi: [10.1097/CIN.0000000000000305](#)]
62. Long AL, Horvath MM, Cozart H, Eckstrand J, Whitehurst J, Ferranti J. Tailoring adverse drug event surveillance to the paediatric inpatient. *Qual Saf Health Care* 2010 Oct;19(5):e40 [FREE Full text] [doi: [10.1136/qshc.2009.032680](#)] [Medline: [20511599](#)]
63. Haffner S, von LN, Wirth S, Thürmann PA. Detecting adverse drug reactions on paediatric wards: intensified surveillance versus computerised screening of laboratory values. *Drug Saf* 2005;28(5):453-464. [Medline: [15853446](#)]
64. Mull HJ, Rosen AK, Shimada SL, Rivard PE, Nordberg B, Long B, et al. Assessing the potential adoption and usefulness of concurrent, action-oriented, electronic adverse drug event triggers designed for the outpatient setting. *EGEMS (Wash DC)* 2015;3(1):1116 [FREE Full text] [doi: [10.13063/2327-9214.1116](#)] [Medline: [25992386](#)]
65. Rosen AK, Mull HJ, Kaafarani H, Nebeker J, Shimada S, Helwig A, et al. Applying trigger tools to detect adverse events associated with outpatient surgery. *J Patient Saf* 2011 Mar;7(1):45-59. [doi: [10.1097/PTS.0b013e31820d164b](#)] [Medline: [21921867](#)]
66. Lemon V, Stockwell DC. Automated detection of adverse events in children. *Pediatr Clin North Am* 2012 Dec;59(6):1269-1278. [doi: [10.1016/j.pcl.2012.08.007](#)] [Medline: [23116524](#)]
67. Patient Safety Organization Privacy Protection. Common Formats for Event Reporting - Hospital Version 2 URL: [https://www.psoppc.org/psoppc\\_web/publicpages/commonFormatsHV2.0](https://www.psoppc.org/psoppc_web/publicpages/commonFormatsHV2.0) [accessed 2018-03-28] [WebCite Cache ID 6yFTMJEMN]

## Abbreviations

**ADE:** adverse drug event  
**AE:** adverse event  
**AHRQ:** Agency for Healthcare Research and Quality  
**DTA:** diagnostic test accuracy  
**EHR:** electronic health record  
**GTT:** Global Trigger Tool  
**HCP:** health care provider  
**IHI:** Institute for Healthcare Improvement  
**MeSH:** medical subject headings  
**NCC MERP:** National Coordinating Council for Medication Error Reporting and Prevention  
**NLP:** natural language processing  
**PPV:** positive predictive value  
**PSI:** Patient Safety Indicator

**QUADAS-2:** Quality Assessment tool for Diagnostic Accuracy Studies-2

*Edited by G Eysenbach; submitted 22.01.18; peer-reviewed by Y Wang, E Berner; comments to author 14.03.18; revised version received 28.03.18; accepted 28.03.18; published 30.05.18*

Please cite as:

Musy SN, Ausserhofer D, Schwendimann R, Rothen HU, Jeitziner MM, Rutjes AWS, Simon M  
Trigger Tool–Based Automated Adverse Event Detection in Electronic Health Records: Systematic Review  
J Med Internet Res 2018;20(5):e198  
URL: <http://www.jmir.org/2018/5/e198/>  
doi: [10.2196/jmir.9901](https://doi.org/10.2196/jmir.9901)  
PMID:

©Sarah N Musy, Dietmar Ausserhofer, René Schwendimann, Hans Ulrich Rothen, Marie-Madlen Jeitziner, Anne WS Rutjes, Michael Simon. Originally published in the Journal of Medical Internet Research (<http://www.jmir.org>), 30.05.2018. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on <http://www.jmir.org/>, as well as this copyright and license information must be included.

## Multimedia Appendix: Hausner et al.'s approach

This is a Multimedia Appendix to a full manuscript published in the J Med Internet Res. For full copyright and citation information see <http://dx.doi.org/10.2196/jmir.9901>.

From our personal bibliographic archive and references of systematic reviews, we located 62 papers for the trigger tool methodology. We randomly divided those papers into two sets of 31: (1) Development set and; (2) Validation set. From the development set, we extracted the frequency of free text terms for title and abstract fields and subject headings (MeSH) with the PubMed PubReMiner tool (a free web service for searches in MEDLINE). This process identified potential candidate terms for inclusion in the search strategy. The search strategy's success was indicated by the number of results it returned and the sensitivity of the development and validation sets. The sensitivity is calculated by the number of papers found with the search strategy divided by the total number of papers in the set. For example, with the search strategy "trigger[tiab] OR triggers[tiab]", we found 31 papers out of 31 of the development set among the 99,429, meaning that it has a sensitivity of 100%. However, a high sensitivity is not enough: the number of results should be low to have a precise strategy. Sensitivities of the development and validation sets both should be around 90% with a low number of results. The table below shows the various steps with the final result (conducted on October 21<sup>st</sup>, 2015).

Search Strategy	Results	Sensitivity development set	Sensitivity validation set
trigger[tiab] OR triggers[tiab]	99,429	100% (31/31)	100% (31/31)
(tool[tiab] OR tools[tiab]) AND (trigger[tiab] OR triggers[tiab])	2,637	100% (31/31)	93.6% (29/31)
adverse[tiab] AND identif*[tiab]	50,525	80.7% (25/31)	87.1% (27/31)
adverse[tiab] AND (trigger[tiab] OR triggers[tiab] OR safety[tiab] OR medical errors[mh])	68,266	90.3% (28/31)	93.6% (29/31)
adverse[tiab] AND (trigger[tiab] OR triggers[tiab] OR detect*[tiab] OR medical errors[mh])	28,943	90.3% (28/31)	93.6% (29/31)
adverse[tiab] AND (trigger[tiab] OR triggers[tiab] OR record[tiab] OR records[tiab] OR medical errors[mh])	14,154	90.3% (28/31)	93.6% (29/31)
adverse[tiab] AND (trigger[tiab] OR triggers[tiab] OR chart[tiab] OR charts[tiab] OR medical errors[mh])	8,802	90.3% (28/31)	93.6% (29/31)
(adverse[tiab] OR medical errors[mh]) AND (detect*[tiab] OR trigger[tiab] OR triggers[tiab]) AND (chart[tiab] OR charts[tiab] OR record[tiab] OR records[tiab])	1,681	74.2% (23/31)	83.9% (26/31)
(adverse[tiab] OR trigger[tiab] OR triggers[tiab]) AND (detect*[tiab] OR identif*[tiab]) AND	6,157	87.1% (27/31)	93.6% (29/31)



(medical errors[mh] OR chart[tiab] OR  
charts[tiab] OR record[tiab] OR records[tiab])

(trigger[tiab] OR triggers[tiab]) AND (adverse[tiab] OR record[tiab] OR records[tiab]) AND (chart[tiab] OR charts[tiab] OR medical errors[mh])	191	77.4% (24/31)	83.9% (26/31)
---------------------------------------------------------------------------------------------------------------------------------------------------------	-----	---------------	---------------

(trigger[tiab] OR triggers[tiab]) AND (chart[tiab] OR charts[tiab] OR identif*[tiab] OR record[tiab] OR records[tiab]) AND (adverse[tiab] OR medical errors[mh])	561	90.3% (28/31)	93.6% (29/31)
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-----	---------------	---------------

## Multimedia Appendix: QUADAS-2

This is a Multimedia Appendix to a full manuscript published in the J Med Internet Res. For full copyright and citation information see <http://dx.doi.org/10.2196/jmir.9901>.

**Phase 1: State the review question: To determine the sensitivity and specificity of the Global Trigger Tool in detecting adverse events in adult inpatients with an electronic/automatic approach, using the manual GTT application as the reference standard.**

<i>Patients (setting, intended use of index test, presentation, prior testing):</i> <b><i>All patients hospitalized for at least 48 hours (inpatient) with any specific disease</i></b>
<i>Index test(s):</i> <b><i>Global Trigger Tool or a modified version (added/removed/modified triggers) used in a semi- or fully automatic way</i></b>
<i>Comparator test:</i> <b><i>None</i></b>
<i>Reference standard and target condition:</i> <b><i>Manual Global Trigger Tool or a modified version (added/removed/modified triggers)</i></b>

## Flow diagram of the study

## DOMAIN 1: PATIENT SELECTION

### A. *Risk of bias*: Could the selection of patients have introduced bias?

*Describe methods of patient selection:*

**1. Was a consecutive or random sample of patients enrolled? Yes / No / Unclear<sup>\*</sup>**

If all accessible patients were selected as a sample or if the process of sampling was done with the method of random sampling, the question will be answered as “yes”.

**2. Was a case-control design avoided? Yes / No / Unclear<sup>\*</sup>**

We excluded case-control studies, as these were deemed inappropriate; therefore, this question will not be used.

**3. Did the study avoid inappropriate exclusions? Yes / No / Unclear<sup>\*</sup>**

This question will be answered with “no” if patients are excluded because of conditions known to be associated with false negative or false positive results. In these situations, automated trigger tool may fail (over/underdiagnose adverse events). Excluding such patient record forms may artificially inflate estimates of diagnostic test accuracy.

**RISK: LOW / HIGH / UNCLEAR**

### B. *Applicability*: Are there concerns that the included patients and setting do not match the review question?

*Describe included patients (advanced vs early disease, symptoms, setting, prior testing, presence of alternative conditions, demographic features, intended use of the test, inpatient duration):*

If a study did not meet the patient population as described in the objective there will be a high concern regarding its applicability.

**CONCERN: LOW / HIGH / UNCLEAR**

<sup>\*</sup> Refers to signaling questions of the original QUADAS-2 tool

<sup>†</sup> Refers to signaling questions added to the original QUADAS-2 tool

## DOMAIN 2: INDEX TEST

### A. **Risk of Bias:** Could the conducting or interpretation of the index test have introduced bias?

Describe the index test and how it was conducted and interpreted:

Description of the preventability

**1. Were the index test results interpreted without knowledge of the results of the reference standard? Yes / No / Unclear<sup>\*</sup>**

Less important for objective tests or if index test is interpreted prior to reference standard.

**2. Were the adverse events defined based on the U.S Food & Drug Administration (FDA) definition and/or on the Institute for Healthcare Improvement (IHI) definition? Yes / No / Unclear<sup>†</sup>**

FDA definition: "Adverse event means any untoward medical occurrence associated with the medical intervention(s) at the hospital, whether or not considered drug related." (<https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?fr=312.32>)

**3. Was the assessment of severity of the adverse events based on the National Coordinating Council for Medication Error Reporting and Prevention (NCC MERP)? Yes / No / Unclear<sup>†</sup>**

- FDA definition: "Adverse event means any untoward medical occurrence associated with the medical intervention(s) at the hospital, whether or not considered drug related." (<https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?fr=312.32>)
- IHI definition: "Any noxious or unintended event occurring in association with medical care." (Griffin FA, Resar RK. IHI Global Trigger Tool for measuring adverse events. Institute for Healthcare Improvement Innovation Series White Paper. 2009.)

**4. Did the development of the algorithm involve clinician(s) and was it based on a test set or an empirical approach? Yes / No / Unclear<sup>†</sup>**

Involvement of a clinician is assumed/shown to improve validity as well as an empirical development approach (<https://www.ncbi.nlm.nih.gov/pubmed/18487779>; <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2909812/>).

**RISK: LOW / HIGH / UNCLEAR**

### B. **Applicability:** Are there concerns that the index test, its conducting, or its interpretation differ from the review question?

<sup>\*</sup> Refers to signaling questions of the original QUADAS-2 tool

<sup>†</sup> Refers to signaling questions added to the original QUADAS-2 tool

If test conducting, technology, setting or interpretation differ from your review question, the results may not be applicable. Example: a higher ultrasound transducer frequency has been shown to improve sensitivity for the evaluation of patients with abdominal trauma.

<b>CONCERN: LOW / HIGH / UNCLEAR</b>
--------------------------------------

\* Refers to signaling questions of the original QUADAS-2 tool

† Refers to signaling questions added to the original QUADAS-2 tool



### DOMAIN 3: REFERENCE STANDARD

**A. Risk of Bias:** *Could the reference standard, its conduct, or its interpretation have introduced bias?*

*Describe the reference standard and how it was conducted and interpreted:*

**1. Is the reference standard likely to correctly classify the target condition? Yes / No / Unclear<sup>\*</sup>**

We are aware that there is no globally accepted reference standard in this field but reason that the full manual chart review and the manual version of the GTT are at the moment the most sensitive methods to detect AEs, although we lack empirical data on the misclassification rates of the full manual chart review and the manual version of the GTT. In this situation, we judged yes if a full manual chart review or a manual version of the GTT were used, and no evidence was found in the study report that was suggestive of potential sources of bias related to the use/development of the full manual chart review and the manual version of the GTT.

**2. Were the reference standard results interpreted without knowledge of the results of the index test? Yes / No / Unclear<sup>\*</sup>**

This question will be scored as “yes” if it is explicitly stated that the adverse event detection occurs without knowledge of the automatic GTT results. The reference standard should be interpreted blind to index test results. Related to degree of subjectivity and order of tests.

**3. Is the inter-rater reliability clearly stated and sufficiently high? Yes / No / Unclear<sup>†</sup>**

Although a clinical diagnostic test accuracy study should be conducted after reliability is more or less established, in the field of GTT, the evidence on reliability is variable. For this reason, we added a signalling question regarding reliability. We classified “yes” if the inter-rater reliability was assessed with sound methods, and was judged to be acceptably high.

**RISK: LOW / HIGH / UNCLEAR**

**B. Applicability:** *Are there concerns that the target condition as defined by the reference standard does not match the question?*

Reference standard outcomes are decisive: if the reference standard does not detect the target condition defined in the review question, results may not be applicable. It is critical to choose a valid/optimal reference standard.

**CONCERN: LOW / HIGH / UNCLEAR**

<sup>\*</sup> Refers to signaling questions of the original QUADAS-2 tool

<sup>†</sup> Refers to signaling questions added to the original QUADAS-2 tool

## DOMAIN 4: FLOW AND TIMING

### A. Risk of Bias: Could the patient flow have introduced bias?

*Describe any patients who did not receive the index test(s) and/or reference standard or who were excluded from the 2x2 table (refer to flow diagram):*

*Describe the time interval and any interventions between index test(s) and reference standard:*

**1. Was there an appropriate interval between index test and reference standard? Yes / No / Unclear<sup>\*</sup>**

Delays between tests can cause misclassification due to recovery or progression to more advanced disease. The length of time which may cause such bias will vary between conditions.

This item is not applicable to the comparison between automated versus full manual chart review or the manual version of the GTT, due to the retrospective nature of both tools. We typically scored yes, unless there was evidence that the time period verified by the manual and automated version of the tools differed. In cases where the latter was true, we judged “no.”

**2. Were all patients included in the analysis? Yes / No / Unclear<sup>\*</sup>**

This question will be scored as “yes” if all patients who were recruited into the study were included in the analysis. No is scored if one or more patients are missing from the 2x2 table. As the study prevalence of AEs can be as low as 2%, we judged high risk of bias if 6% or more of the patients were missing from the 2x2 table—assuming that one-third of those missing would not be random, potentially affecting our estimates of the prevalence.

**3. Was the completeness of electronic health records (EHRs) data discussed and addressed? Yes / No / Unclear<sup>†</sup>**

If yes, the risk of bias will be lower, since it takes into account missing data for the analysis.

**RISK: LOW / HIGH / UNCLEAR**

<sup>\*</sup> Refers to signaling questions of the original QUADAS-2 tool

<sup>†</sup> Refers to signaling questions added to the original QUADAS-2 tool

## **Multimedia Appendix: Trigger tool Quality Assessment Tool**

This is a Multimedia Appendix to a full manuscript published in the J Med Internet Res. For full copyright and citation information see <http://dx.doi.org/10.2196/jmir.9901>.

**STATE your review question here, and answer all signalling questions and judgments in light of this review question.**

1. To determine the prevalence of AEs as detected by an electronic/automated or semi-automatic trigger tool in various adult inpatient populations.
2. To describe the reliability of electronic/automated trigger tools
3. To explore methods of phenotyping if adverse event detection in EHRs was used in the international literature.

**P** Patient records of patients hospitalized at least 48 hours (inpatient) with any specific disease admitted to any ward

**I** Global Trigger Tool or a modified version (added/removed/modified triggers) used in an automatic or semi-automatic way

**C** Not applicable

**O** Prevalence overall, by type of AE, and by type of hospital or ward.

## DOMAIN 1A: PATIENT SELECTION

*Describe methods of patient selection:*

**A. Risk of bias:** *Could the selection of patient records have introduced bias?*

**1. Was the participation rate of eligible persons at least 50%? Yes / No / Unclear**

If the rate is less, this likely introduces bias (From <http://www.nhlbi.nih.gov/health-pro/guidelines/in-develop/cardiovascular-risk-reduction/tools/cohort>).

**2. Was a consecutive or random sample of patient records enrolled? Yes / No / Unclear**

Reflect if all the subjects selected or recruited were from the same or similar populations (including the same time period)? Were inclusion and exclusion criteria for being in the study pre-specified and applied uniformly to all participants in a consecutive manner? If all accessible patient records were selected as a sample or if the process of sampling was done with the method of random sampling, this question will be answered as “yes”.

**3. Did the study consider patients covering a broad range of indications for hospitalisation? Yes / No / Unclear**

This question will be answered with “no” when patients with very different profiles are not considered by exclusion from study entry. Such exclusions are highly likely to alter the estimates of prevalence. This is a situation where GTT might over/underestimate adverse events. For example: exclusion of certain groups of patients due to extended lengths of stay or high numbers of transfers.

**RISK: LOW / HIGH / UNCLEAR**

**B. Applicability:** *Are there concerns that the included patients and setting do not match the review question?*

If a study did not meet the patient population as described in the objective there will be a high concern regarding its applicability. In this specific review, we allow for a broad range of settings and study populations.

**CONCERN: LOW / HIGH / UNCLEAR**

## DOMAIN 1B: REVIEWER/ALGORITHM SELECTION

*Describe methods to sample patients to be included in the reliability estimations:*

*State the determined number of reviewers:*

*State the number of patient records included in the reliability estimations:*

*State the number of replicated observations:*

*Describe the reviewer characteristics (e.g., training, experience):*

*Did raters judge/analyse the records independently (e.g., stage 1 screening)?*

**A. Risk of bias:** *Could the selection of reviewers/algorithms have introduced bias?*

**1. Was the selection of the reviewer(s) based on his/her experience and/or professional background in the clinical setting? Yes / No / Unclear**

A lack of experience on the part of the reviewer(s) in the clinical setting may introduce bias. For reviewers with appropriate clinical backgrounds the bias might be lower concerning misclassification rates.

**2. Were the reviewer(s) trained on using and applying trigger tool methodology? Yes / No / Unclear**

A lack of trigger tool training and application knowledge may introduce bias. For reviewer(s) with more training the bias might be lower.

**3. Do the reviewer(s) have experience in applying the trigger tool or another retrospective chart review methodology? Yes / No / Unclear**

A lack of trigger tool experience may introduce bias. For reviewer(s) with more trigger tool experience the bias might be lower.

**4. Did the study use a test and validation sets to develop the algorithm? Yes / No / Unclear**

The development of EHR algorithms always leads to choices (e.g., whether they should be geared towards sensitivity or precision). Using the split-half method is crucial to assess the efficacy of the proposed algorithm.

**5. Is the inter-rater reliability clearly stated and sufficiently high? Yes / No / Unclear**

Although a clinical diagnostic test accuracy study should be conducted after reliability is more or less established, in the field of GTT, the evidence on reliability is variable. For this reason, we added a signalling question regarding reliability. We classified “yes” if the inter-rater reliability was clearly assessed with sound methods, and was judged to be acceptably high.

**RISK: LOW / HIGH / UNCLEAR**

**B. Applicability:** *Are there concerns that the reviewer(s) do not match the review question?*

For example, if the profile(s) of the reviewer(s) applying the trigger tool in the study differ substantially from those of healthcare professionals who would apply it in clinical practice, a high concern may arise.

<b>CONCERN: LOW / HIGH / UNCLEAR</b>
--------------------------------------

## DOMAIN 2: AUTOMATIC DETECTION METHOD

*Description of the method*

### **A. Risk of Bias:** *Could the conducting or interpretation of the index test have introduced bias?*

**1. Was the application of the algorithms fully automatic? Yes / No / Unclear**

Algorithms implemented in a semi-automatic way with different process steps are expected to introduce more bias (semi-automatic processes being more prone to errors).

**2. Did the development of the algorithm involve clinician(s) and was it based on a test set or an empirical approach? Yes / No / Unclear**

Involvement of a clinician is assumed/shown to improve validity as well as an empirical development approach (<https://www.ncbi.nlm.nih.gov/pubmed/18487779>; <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2909812/>).

**3. Was the selection of triggers based on literature review and/or consultation with a group of experts in the field? Yes / No / Unclear**

Empirical evidence or at least a biological rationale should exist for each of the triggers included in the tool. This particularly applies to *new triggers not part of the original IHI set of triggers*.

**4. Are the triggers used the same for all settings and/or the hospitals participating in the study? Yes / No / Unclear**

If differing versions of triggers are in use across participating units, a strong rationale should have been given to adapt them. For example, it is acceptable to adapt triggers according to national circumstances where a specific drug is not licensed.

**5. Was the presence of any adverse event checked/controlled at the admission of the patient to the unit/hospital? Yes / No / Unclear**

If not, the unit/hospital might conclude that those AEs were the consequences of their own care to the patients.

**RISK: LOW / HIGH / UNCLEAR**

### **B. Applicability**

**1. Are there concerns that the trigger tool test, its conducting, or its interpretation differ from the review question?**

If test conducting, technology, setting or interpretation differ from your review question, the results may not be applicable. For example: triggers are not related to the IHI GTT; AEs captured in the study are more triggers than AEs.



**CONCERN: LOW / HIGH / UNCLEAR**

## DOMAIN 3: OUTCOMES

### A. *Risk of Bias*: Could the definition of outcomes have introduced bias?

*Describe the definition(s) of AE:*

*Describe the how prevalence was measured:*

*Describe your judgment, if you deem the definitions to be standard, or deviating from our review definitions and describe if you deem reliability sufficiently high:*

**1. Were the adverse events defined based on the U.S Food & Drug Administration (FDA) definition and/or on the Institute for Healthcare Improvement (IHI) definition? Yes / No / Unclear**

- FDA definition: “Adverse event means any untoward medical occurrence associated with the medical intervention(s) at the hospital, whether or not considered drug related.” (<https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?fr=312.32>)
- IHI definition: “Any noxious or unintended event occurring in association with medical care.” (Griffin FA, Resar RK. IHI Global Trigger Tool for measuring adverse events. Institute for Healthcare Improvement Innovation Series White Paper. 2009.)

**2. Was the assessment of the adverse events’ severity based on the National Coordinating Council for Medication Error Reporting and Prevention (NCC MERP)? Yes / No / Unclear**  
Classification of the AEs’ severity follow: <http://www.nccmerp.org/sites/default/files/indexColor2001-06-12.pdf>

**3. For prevalence, were the outcome measures (dependent variables) clearly defined? Yes / No / Unclear**

A lack of a clear outcome measure definition might lead to different results, thus introducing bias.

**4. For prevalence, were the definitions consistently applied across all study participants? Yes / No / Unclear**

If not, AE prevalence values might differ between patient groups, thus introducing bias.

**RISK: LOW / HIGH / UNCLEAR**

#### DOMAIN 4: FLOW AND TIMING

##### **A. Risk of Bias:** *Could the patient flow have introduced bias?*

- 1. Was the completeness of electronic health records (EHRs) data discussed and addressed? *Yes / No / Unclear***

If yes, the risk of bias may be lower, since it takes into account missing data for the analysis.

- 2. Were all patients included in the analysis of prevalence? *Yes / No / Unclear***

This question will be scored as "yes" if all patients who were recruited into the study were included in the analysis. No is scored if one or more patients are missing. As the study prevalence of AEs can be as low as 2.x%, we judge high risk of bias if 10% or more of the patients are omitted from the AE prevalence calculations. This assumes that one third of the missing would not be at random, potentially affecting our prevalence estimates.

<b><i>RISK: LOW / HIGH / UNCLEAR</i></b>
------------------------------------------